УДК 681.3.01

МЕТОДИКА ОЦЕНКИ ПАРАЛЛЕЛЬНОСТИ ВАРИАНТОВ ТЕМАТИЧЕСКОГО ТЕСТА НА ОСНОВЕ СТАТИСТИЧЕСКИХ МЕТОДОВ

О.В. Марухина, О.Г. Берестнева, Л.И. Рахматуллина

Томский политехнический университет
E-mail:olgmik@osu.cctpu.edu.ru

Представлена методика оценки параллельности вариантов текущего теста по математике студентов всех факультетов Томского политехнического университета на основе математико-статистических методов. Проведен анализ результатов тестирования студентов в 2004 г. и сделаны соответствующие выводы о параллельности вариантов теста.

В современной системе образования независимая аттестация студентов является наиболее объективной оценкой их знаний, потенциала их умственных возможностей. Проведение реформы образования и современная стратегия Томского политехнического университета (ТПУ), а именно – интеграция в международное образовательное пространство и его конкурентоспособность сделало систему тестирования востребованной [1]. В связи с этим в Центре тестирования Томского политехнического университета разработана система независимой оценки качества знаний студентов по общеобразовательным дисциплинам. Контрольно-измерительные материалы по дисциплинам представлены в нескольких вариантах. Например, по математике, имеется двадцать один вариант тестовых заданий. Из этого вытекает проблема параллельности ("одинаковости") этих вариантов тестовых заданий, и, как следствие, качество оценки знаний студентов, и ее объективность.

Таким образом, объектом исследования являются студенты ТПУ, участвующие в процессе текущего контроля и оценки качества знаний по математике, а предметом исследования обозначим варианты тестовых заданий (в данном исследовании – тест по высшей математике). Всего в тестировании принимали участие 1001 человек – студенты-первокурсники технических специальностей ТПУ.

В табл. 1 приведены результаты первичной обработки данных тестирования по двадцати одному варианту теста по математике. Оценка трудности каждого варианта δ_j проводилась по алгоритму, описанному в [2]. В работе использовалось разработанное авторами программное обеспечение для оценки тестов LogitModels [3].

Таблица 1. *Результаты первичной обработки данных тестирования*

Вариант теста	Количество испытуемых	Не справились с тестом	Справились со всеми заданиями теста	Трудность теста, δ_j
B1	49	0	0	−0,01
B2	50	0	0	−0,04
B3	59	0	0	0,12
B4	52	0	0	0
B5	53	1	0	0,22
B6	42	0	0	0,15
B7	46	0	0	0,04
B8	41	2	0	0,23
B9	42	1	0	0,06
B10	48	1	0	0,10
B11	50	1	0	0,09
B12	38	0	0	0,10
B13	45	0	0	0,03
B14	48	0	0	0,39
B15	56	2	0	0,11
B16	43	0	0	−0,05
B17	45	1	0	0,24
B18	53	0	0	0,02
B19	43	0	0	0,38
B20	50	0	1	−0,12
B21	48	0	0	0,02

На рис. 1 приведен график изменения трудности вариантов теста по математике.

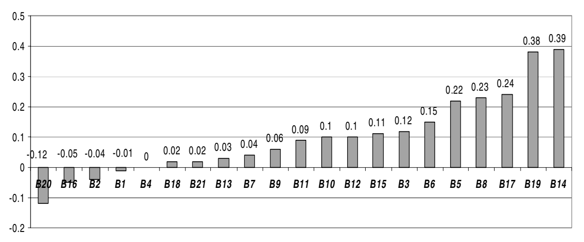


Рис. 1. *Геометрический профиль трудности вариантов теста по математике. По оси абсцисс – номера вариантов теста; по оси ординат – логиты трудности*

Для оценки значимости разброса значений трудности вариантов теста был использован коэффициент вариации (1):

$$Cv = \frac{S_x}{\bar{x}} \cdot 100\ \%, \qquad (1)$$

где S_x – стандартное отклонение распределения трудности вариантов теста, $\bar{x}$ – среднее значение трудности вариантов теста. Различные признаки характеризуются различными коэффициентами вариации. Но в отношении одного и того же признака значение этого показателя C_v остаётся более или менее устойчивым и при симметричных распределениях обычно не превышает 50 %. При сильно асимметричных рядах распределения коэффициент вариации может достигать 100 % и даже выше. Для рассматриваемого теста коэффициент составит:

$$Cv = \frac{S_x}{\bar{x}} \cdot 100\ \% = (0{,}099/0{,}132) \cdot 100\ \% = 134\ \%.$$

Варьирование считается значительным при C_v 25 %, то есть изучаемая совокупность считается разнородной, следовательно, по трудности варианты теста **непараллельны**.

Оценка однородности вариантов

Для оценки связи между результатами выполнения двух заданий теста или вариантов теста была использована формулу коэффициента корреляции (2):

$$\varphi_{jl} = \frac{p_{jl} - p_j \cdot p_l}{\sqrt{p_j \cdot q_j \cdot p_l \cdot q_l}}, \qquad (2)$$

где j,l – номера заданий теста, p_{jl} – доля испытуемых, выполнивших правильно оба задания теста, т.е. доля тех, кто получил один балл по обоим заданиям; p_j и p_l – доля испытуемых, правильно выполнивших j-ое и l-ое задание; q_j и q_l – доля испытуемых, неправильно выполнивших j-ое и l-ое задание; q_i=1–p_j; q_l=1–p_l.

Матрица стандартизированных значений трудности заданий теста по вариантам приведена в табл. 2.

Таблица 2. *Стандартизированные значения трудности заданий теста A_i по вариантам B_j*

	A1	*A2*	*A3*	*A4*	*A5*	*A6*	*A7*	*A8*	*A9*	*A10*	*A11*
B1	−1,41	−1,53	0,18	0,38	−0,40	−0,21	1,55	0,18	0,38	0,71	0,60
B2	−1,32	0,97	−1,57	0,97	0,62	−0,89	0	−0,69	0,73	0,73	0,73
B3	−0,94	−1,18	−0,31	−0,31	−0,23	−0,39	−0,55	1,54	1,38	2,14	0,45
B4	−1,58	−1,70	0,36	0,68	0,36	−1,83	0,26	0,79	0,26	0,91	2,27
B5	−1,73	−1,07	−0,16	−0,25	−0,51	−0,25	1,28	−0,07	3,01	0,30	1,28
B6	−0,52	−1,65	−0,30	−0,30	−0,86	0,15	1,06	0,04	−0,41	2,32	1,79
B7	−0,88	0,61	0,03	−0,38	−0,38	−1,40	−0,18	−0,18	1,81	1,38	0,61
B8	−0,19	0,58	−0,06	0,44	0,18	−1,31	0,31	−1,04	1,80	0,06	1,02
B9	1,11	0,29	−1,87	−0,57	0,58	0,43	0,58	−1,36	1,57	−0,22	1,11
B10	−0,21	−0,01	−2,13	−0,88	0,31	0,31	0,67	0,54	−0,21	1,44	1,64
B11	0,12	−1,28	−0,64	−0,84	1,63	−0,07	0,52	0,22	0,32	0,41	0,32
B12	−1,72	−1,40	−1,89	1,18	0,27	1,37	−0,36	0,69	−0,11	0,27	3,00
B13	−0,40	0,23	−0,40	−1,15	0,34	−1,40	0,23	0,02	2,07	−0,30	1,43
B14	−2,49	−1,75	−0,05	0,77	0,77	−0,76	0,39	−0,15	4,07	0,39	2,03
B15	−2,18	−0,50	0,14	0,06	−0,42	0,23	0,81	0,14	−0,02	1,64	1,47
B16	1,07	−0,79	0,60	−0,25	0,33	−2,73	0,90	−0,47	0,46	0,75	0,90
B17	−1,62	2,60	−0,69	−0,49	0,88	−1,09	0,71	0,05	0,05	1,49	0,88
B18	0,86	−0,55	0,16	−0,82	0,75	−2,18	0,16	−0,37	0,25	0,98	1,37
B19	−2,09	−1,46	−0,82	−0,22	2,76	−1,19	0,55	−0,10	1,00	3,30	1,00
B20	−2,14	−3,76	0,40	−0,07	1,03	−0,46	1,39	0,40	0,12	0,60	0,50
B21	0,43	−1,34	0,02	1,05	0,12	−0,64	1,73	−1,02	−0,36	0,12	0,54

Значения коэффициента корреляции Спирмена между результатами по отдельным вариантам теста сводятся в матрицу (табл. 3). В табл. 3 выделены значения коэффициентов корреляции, уровень значимости p для которых меньше 0,05 (эти значения коэффициентов корреляции оказались статистически значимыми на 5 % уровне. Все расчеты производились с использованием пакета Statistica 6.0). Это означает, что для данных пар вариантов су-

Таблица 3. *Матрица коэффициентов корреляции между вариантами теста*

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20	B21
B1	1,00																				
B2	0,20	1,00																			
B3	0,50	0,14	1,00																		
B4	0,58	0,24	**0,76**	1,00																	
B5	**0,85**	0,14	**0,61**	0,46	1,00																
B6	**0,81**	-0,07	0,47	0,52	**0,64**	1,00															
B7	0,42	0,47	0,47	0,38	**0,63**	0,17	1,00														
B8	0,31	**0,76**	0,05	0,18	0,40	-0,15	**0,64**	1,00													
B9	0,04	0,12	-0,11	-0,24	0,19	-0,12	0,09	0,45	1,00												
B10	0,46	0,16	0,37	0,41	0,42	**0,66**	0,20	0,02	0,22	1,00											
B11	0,48	-0,07	0,47	0,34	0,47	0,34	0,16	0,10	0,45	**0,62**	1,00										
B12	0,30	0,33	0,44	0,41	0,26	0,50	-0,14	0,03	0,06	0,51	0,15	1,00									
B13	0,20	0,38	0,30	0,23	0,49	-0,12	**0,63**	**0,71**	0,57	0,37	0,55	-0,01	1,00								
B14	**0,62**	0,46	0,58	0,55	**0,68**	0,24	0,45	**0,68**	0,32	0,17	0,49	0,40	0,58	1,00							
B15	**0,80**	-0,01	0,49	0,53	**0,67**	**0,97**	0,26	-0,07	-0,19	**0,66**	0,35	0,48	-0,04	0,31	1,00						
B16	0,43	-0,22	0,04	0,30	0,27	0,28	0,18	0,18	0,43	0,20	0,46	-0,31	0,24	0,19	0,18	1,00					
B17	0,20	**0,67**	0,24	0,34	0,22	0,07	**0,62**	0,53	0,04	0,56	0,33	0,10	**0,62**	0,30	0,22	-0,04	1,00				
B18	0,30	-0,01	0,40	0,50	0,29	0,23	0,41	0,26	0,48	0,42	**0,63**	-0,08	0,52	0,35	0,18	**0,80**	0,30	1,00			
B19	**0,64**	0,33	**0,79**	**0,66**	**0,64**	0,43	0,49	0,35	0,17	0,55	**0,80**	0,36	0,56	**0,78**	0,51	0,22	0,56	0,54	1,00		
B20	**0,67**	-0,06	0,50	0,57	0,56	0,51	0,21	0,09	0,02	0,56	**0,84**	0,15	0,40	0,54	0,59	0,41	0,38	0,47	**0,81**	1,00	
B21	0,58	0,07	-0,06	0,34	0,20	0,36	-0,17	0,24	0,26	0,15	0,38	0,11	0,00	0,40	0,27	**0,70**	-0,08	0,38	0,26	0,47	1,00

ществует сильная положительная связь, т.е. их можно считать параллельными. Однако, далеко не все варианты связаны такой связью. Это подтверждает вывод о **непараллельности** вариантов теста в целом.

Определение параллельности на основе кластерного анализа

Для определения однородных по трудности групп вариантов теста по математике был использован кластерный анализ (методы Варда и k-средних). Кластерный анализ предназначен для разбиения множества объектов на заданное или неизвестное число классов на основании некоторого математического критерия качества классификации (*cluster* (англ.) – гроздь, пучок, скопление, группа элементов, характеризуемых каким-либо общим свойством). Критерий качества классификации в той или иной мере отражает следующие неформальные требования:

1) внутри групп объекты должны быть тесно связаны между собой;
2) объекты разных групп должны быть далеки друг от друга;
3) при прочих равных условиях распределения объектов по группам должны быть равномерными.

Требования 1 и 2 выражают стандартную концепцию компактности классов разбиения; требование 3 состоит в том, чтобы критерий не навязывал объединения отдельных групп объектов.

Многие процедуры при кластеризации совершаются ступенчато. Это означает, что два наиболее близко расположенных объекта x_i и x_j объединяются и рассматриваются как один кластер. Это приводит к тому, что число объектов уменьшается и становится равным n–1, причем один кластер будет содержать два объекта, а остальные по одному. Процесс можно повторять до тех пор, пока все объекты не сгруппируются в один кластер. Наиболее подходящее разбиение выбирает чаще всего сам исследователь, которому предоставляется дендрограмма, отображающая результаты группирования объектов на всех шагах алгоритма кластеризации.

Традиционно различают классификации иерархические и неиерархические (называемые иногда структурными). Соответственно можно разделить алгоритмы получения этих классификаций.

Принцип работы иерархических алгоритмов состоит в последовательном объединении в кластер сначала самых близких, а затем и всё более отдалённых друг от друга элементов. Большинство из этих алгоритмов исходит из матрицы сходства (расстояний), и каждый отдельный элемент рассматривается вначале как отдельный кластер. Общая схема такой иерархической группировки может быть представлена как повторяющееся приложение трех операций к мерам расстояния объект (кластер) – объект (кластер):

1) найти наименьшее расстояние d_{S_1,S_2} между объектом (кластером) S_1 и объектом (кластером) S_2;
2) объединить S_1 и S_2 в один кластер, присвоив общий индекс $S_1 \bigcup S_2$;
3) вычислить расстояние $d_{S,S_1 \bigcup S_2}$ от кластера $S_1 \bigcup S_2$ до любого другого объекта (кластера) S.

Результаты кластеризации вариантов теста по трудности представлены на рис. 2 в виде иерархической дендрограммы. Как видно из рис. 2, все варианты теста можно разбить на три кластера. Для уточнения результата были проведена кластеризация по методу k-средних.

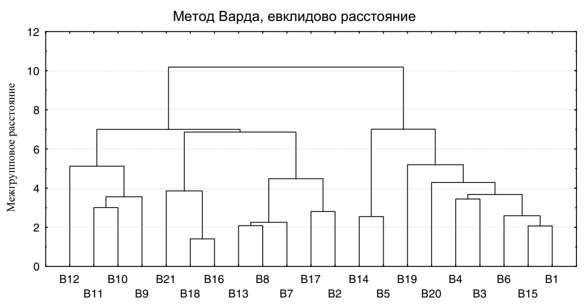


Рис. 2. *Иерархическая дендрограмма результатов кластеризации*

В табл. 4 представлен результат кластеризации по методу k-средних (группировка вариантов теста по трудности).

Таблица 4. *Результат кластеризации по методу k-средних*

Кластер 1	***B2***	***B7***	***B8***	***B9***	***B11***	***B13***	***B16***	***B17***	***B18***	***B21***
Расстояние до центра кластера	0,70	0,62	0,46	0,85	0,74	0,59	0,75	0,98	0,62	0,86
Кластер 2	***B1***	***B3***	***B6***	***B10***	***B12***	***B15***				
Расстояние до центра кластера	0,60	0,74	0,54	0,69	0,91	0,47				
Кластер 3	***B4***	***B5***	***B14***	***B19***	***B20***					
Расстояние до центра кластера	0,69	0,75	0,82	0,96	0,84					

Число искомых кластеров задавалось равным 3. В первом столбце табл. 5 приведен список переменных (заданий теста), далее идут суммы квадратов (SS) и степени свободы (df), затем F-критерий Фишера и в последнем столбце – достигнутый уровень значимости p.

Таблица 5. *Результаты дисперсионного анализа*

Задания	Сумма квадратов *SS*	Степень свободы *df*	Сумма квадратов *SS*	Степень свободы *df*	*F*-критерий Фишера	Достигнутый уровень значимости *p*
A1	13,17	2	11,88	18	9,98	0,00
A2	15,42	2	19,24	18	7,22	0,00
A3	1,21	2	11,67	18	0,93	0,41
A4	0,54	2	8,80	18	0,55	0,59
A5	3,59	2	9,35	18	3,45	0,05
A6	7,36	2	11,17	18	5,93	0,01
A7	0,27	2	7,14	18	0,34	0,71
A8	4,09	2	4,78	18	7,71	0,00
A9	6,33	2	21,23	18	2,68	0,10
A10	3,11	2	13,06	18	2,14	0,15
A11	1,70	2	7,54	18	2,02	0,16

Табл. 5 дисперсионного анализа результатов кластеризации на три кластера показывает необходимость отклонения нулевой гипотезы о равенстве групповых средних по 5 заданиям из 11 (для которых достигнутый уровень значимости оказался более 5 %).

Ниже приведен график (рис. 3) средних значений всех переменных по отдельным кластерам. В табл. 6 приведены соответствующие числовые значения.

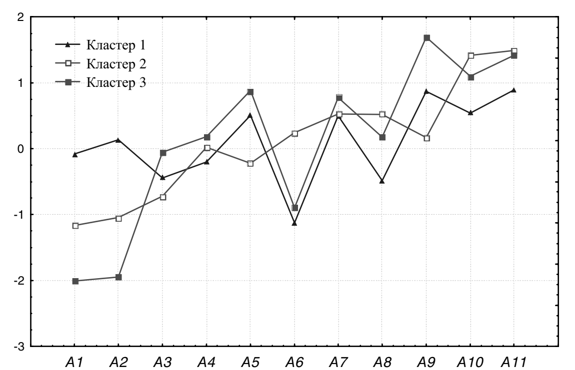


Рис. 3. *Графики средних значений каждого кластера по переменным (заданиям)*

Таблица 6. *Средние значения переменных в каждом кластере*

Переменные	**Значение центроидов**		
	Кластер 1	**Кластер 2**	**Кластер 3**
A1	-0,08	-1,16	-2,01
A2	0,13	-1,05	-1,95
A3	-0,44	-0,72	-0,05
A4	-0,20	0,02	0,18
A5	0,51	-0,22	0,88
A6	-1,13	0,24	-0,90
A7	0,50	0,53	0,77
A8	-0,48	0,52	0,17
A9	0,87	0,17	1,69
A10	0,54	1,42	1,10
A11	0,89	1,49	1,42

Результат проведения кластерного анализа указал на различие (непараллельность) вариантов теста по математике. Сравнивая средние значения трудности каждого кластера по заданиям, следует отметить, что задания *A*1, *A*2, *A*4, *A*5, *A*6 и *A*8 наиболее разнородны, что хорошо видно и на графике средних значений трудности каждого кластера.

Таким образом, кластерный анализ определил три кластера, где внутри каждого кластера варианты являются параллельными, а между собой эти группы по трудности непараллельны. Рекомендуется ввести две различные системы шкалирования результатов тестирования по математике для двух кластеров.

В результате проведенных исследований авторами статьи разработана методика оценки параллельности вариантов теста, которая включает в себя следующие этапы:

1. Расчет стандартных значений уровня трудности заданий по двадцати одному вариантам с использованием специализированного программного обеспечения LogitModels [3, 4];
2. Систематизацию результатов стандартных значений уровня трудности в виде матриц средних значений уровня трудности по заданиям тестов или по вариантам теста;
3. Проверку параллельности вариантов теста по критериям:
   - коэффициент вариации средних стандартных значений уровня трудности вариантов, где сравнительным критерием является значение коэффициента вариации (должно быть $C_v < 25$ %);
   - корреляционный анализ средних стандартных значений уровня трудности заданий и средних стандартных значений уровня трудности вариантов, где сравнительным критерием является коэффициент корреляции (должен быть $r_j \to 1$), что говорит о сильной положительной связи вариантов;
   - кластерный анализ (должна быть одна группа вариантов).
4. Вывод о параллельности.

Структурная схема разработанной методики представлена на рис. 4.



Рис. 4. *Структурная схема разработанной методики оценки параллельности вариантов теста*

Проверка параллельности вариантов по критериям, рассчитываемым в разработанной методике, показала, что все три критерия указывают на принятие нулевой гипотезы о непараллельности вариантов теста.

Работа частично поддержана РФФИ (проект № 03-06-80128

СПИСОК ЛИТЕРАТУРЫ

1. Берестнева О.Г., Иванкина Л.И., Марухина О.В., Пермяков О.Е. Концепция качества образования в техническом вузе // Качество образования: системы управления, достижения, проблемы: Матер. V Междунар. научно-метод. конф. – Новосибирск: Изд-во НГТУ, 2003. – Т. 1. – С. 64–68.
2. Челышкова М.Б. Теория и практика конструирования педагогических тестов: Учебное пособие. – М.: Логос, 2002. – 432 с.
3. Марухина О.В. Алгоритмы обработки информации в задачах оценивания качества обучения студентов вуза на основе экспертно-статистических методов: Дис. ... канд. техн. наук: 05.13.01. – Томск, 2003. – 165 с.
4. Берестнева О.Г., Марухина О.В. Методы многомерного анализа данных в задачах оценки качества образования // Радиоэлектроника. Информатика. Управление. – 2002. – № 1. – С. 15–26.
5. Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. – М.: Прометей, 2000. – 168 с.